

# LINGUISTICS WITH CLARIN TTNWW & PAQU

Jan Odijk

LOT Winterschool

Amsterdam, 2015-01-15

# OVERVIEW

- Enrich your own resources (TTNWW)
- Enrich your own resources (other applications)
- Search in your own enriched resources (AutoSearch)
- Search in your own enriched resources (PaQu)
- Conclusions

# OVERVIEW

- **Enrich your own resources (TTNWW)**
  - Enrich your own resources (other applications)
  - Search in your own enriched resources (AutoSearch)
  - Search in your own enriched resources (PaQu)
  - Conclusions

- TTNWW
  - *TST tools voor het Nederlands als Webservices in een Workflow* (LST Tools for Dutch as web services in a work flow)
  - Goal: to make existing components for Dutch from previous projects (*inter alia* CGN and STEVIN) available as web services and combine them into work flows for use by HSS researchers with little or no technical background.

- Web service: software that can exchange information with other software over the internet
- (Web) application: software targeted at a specific user group with a dedicated (web) interface
- Work Flow: combinations of web services

- **Work Flow** describes the tasks, procedural steps, organizations or people involved, required input and output information, and tools needed for each step in a business process

- **Work Flow** describes the tasks, procedural steps, ~~organizations or people involved~~, required input and output information, and tools needed for each step in a business process
- (sometimes called *processing chain* in the context of software)

# EXISTING COMPONENTS

- **Text**

- Orthographic Normalisation ([TICCLops](#), UvT)
- Part of Speech tagging ([Frog](#), UvT/RUN)
- Parsing ([Alpino](#), RUG)
- Named Entity Recognition ([Frog](#), Ghent)
- Semantic role labelling (MBSRL, UU)
- Coreference Assignment ([COREA](#), Antwerpen/Ghent)

- **Speech**

- (Partial) transcription of speech ([SPRAAK](#), KUL)



# EXISTING COMPONENTS

- For
  - Enrichment of text with corrections of OCR-input
  - Enrichment of text with linguistic annotations
  - Indexation of audio-visual files to make them searchable

# EXISTING COMPONENTS

- Before TTNWW
  - These required local installation (most under Linux only)
  - Required running by the developers
- After TTNWW
  - All available via the TTNWW web application
  - Most available individually as a web service / application

**Frog**  
 mijnProject

---

**Status**

Accepting new input files and selection of parameters Abort and delete project

---

**Input**

**Input files**

Show  entries Search:

Input File	Template	Format	Actions
No data available in table			

Showing 0 to 0 of 0 entries First Previous Next Last

---

**Upload a file from disk**

Use this to upload files from your computer to the system.

**Step 1** First select what type of file you want to add:

**Step 2** Set the parameters for this type of file:  
Select a type first

**Step 3**

---

**Grab a file from the web**

Retrieves an input file from another location on the web.

**Step 1** First select the desired input type:

**Step 2** Set the parameters for this type of file:  
Select a type first

**Step 3** Enter the URL where to retrieve the file

**Step 4**

---

**Add input from browser**

You can create and add new files from within your browser:

---

**Parameter Selection**

**Modules**

**Skip modules**  
Are there any components you want to skip? Skipping the parser and chunker speeds up the process

# EXISTING COMPONENTS

- Before TTNWW
  - Were command line tools
  - Working on a single system
  - No web services
- CLAM
  - Computational Linguistics Application Mediator
  - Generic wrapper to turn existing software easily into a web service
  - Most TTNWW components use CLAM

# EXISTING COMPONENTS

- **Work flows (simplified examples)**

- Part of Speech tagging:

- Tokenization -> sentence splitting -> Pos tagging

- Parsing

- Tokenization -> sentence splitting -> pos-tagging -> parsing

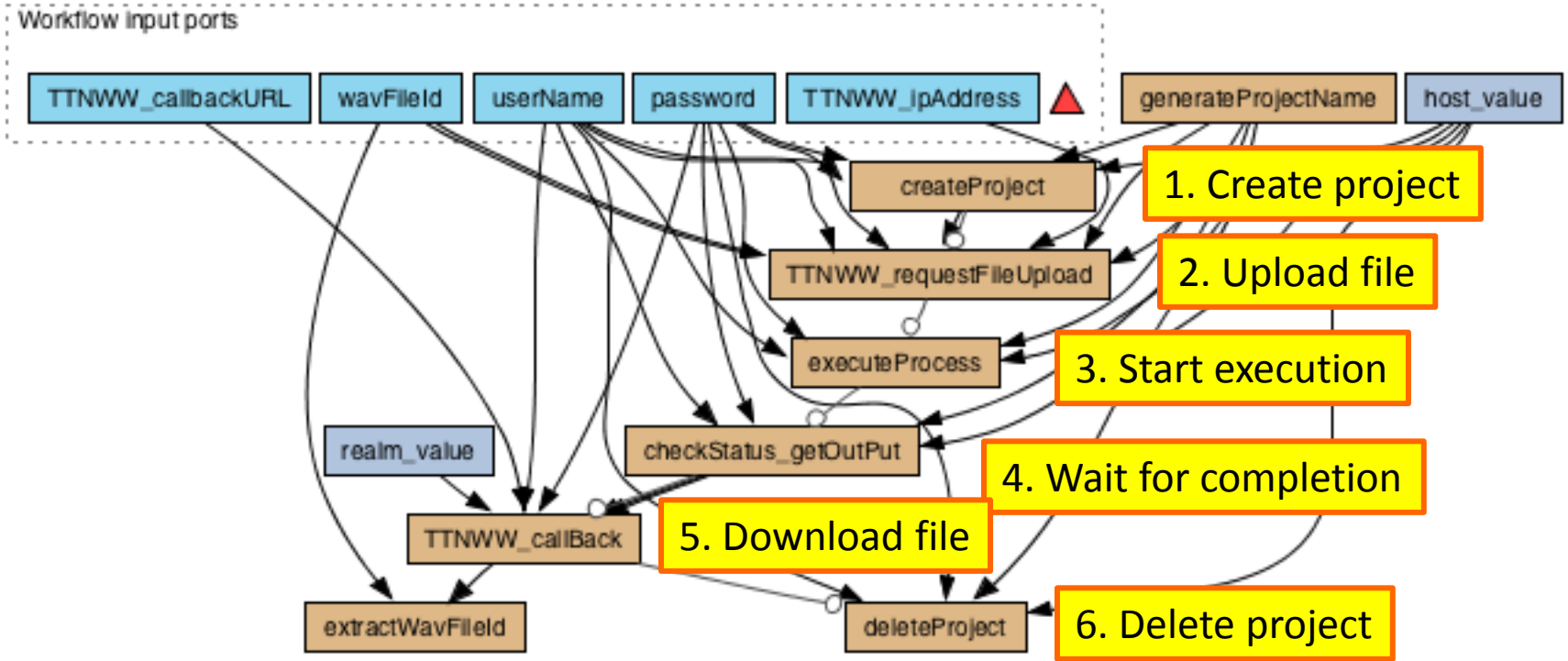
- Named Entity Recognition

- Pos-tagging -> NER

- Semantic role labelling (MBSRL, UU)

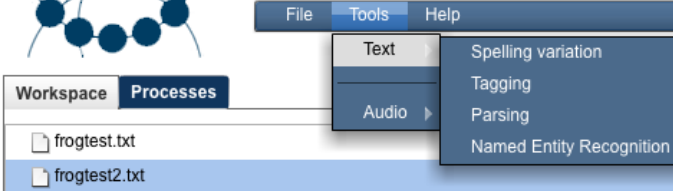
- Parsing -> semantic role labeling

# EXISTING COMPONENTS

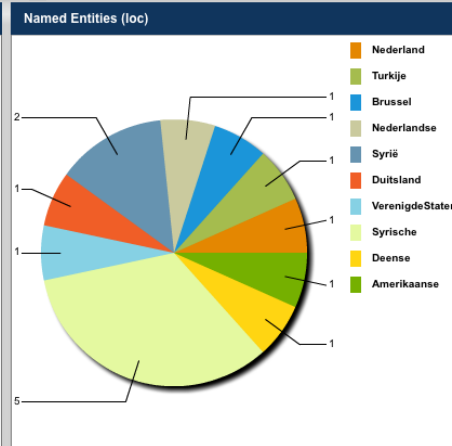
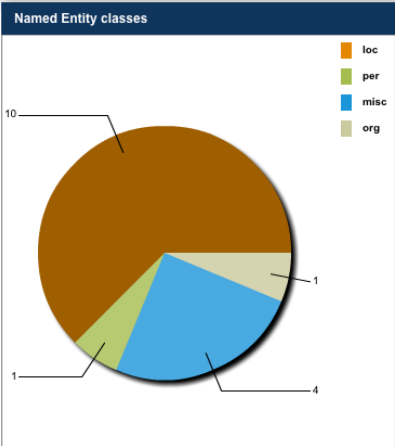
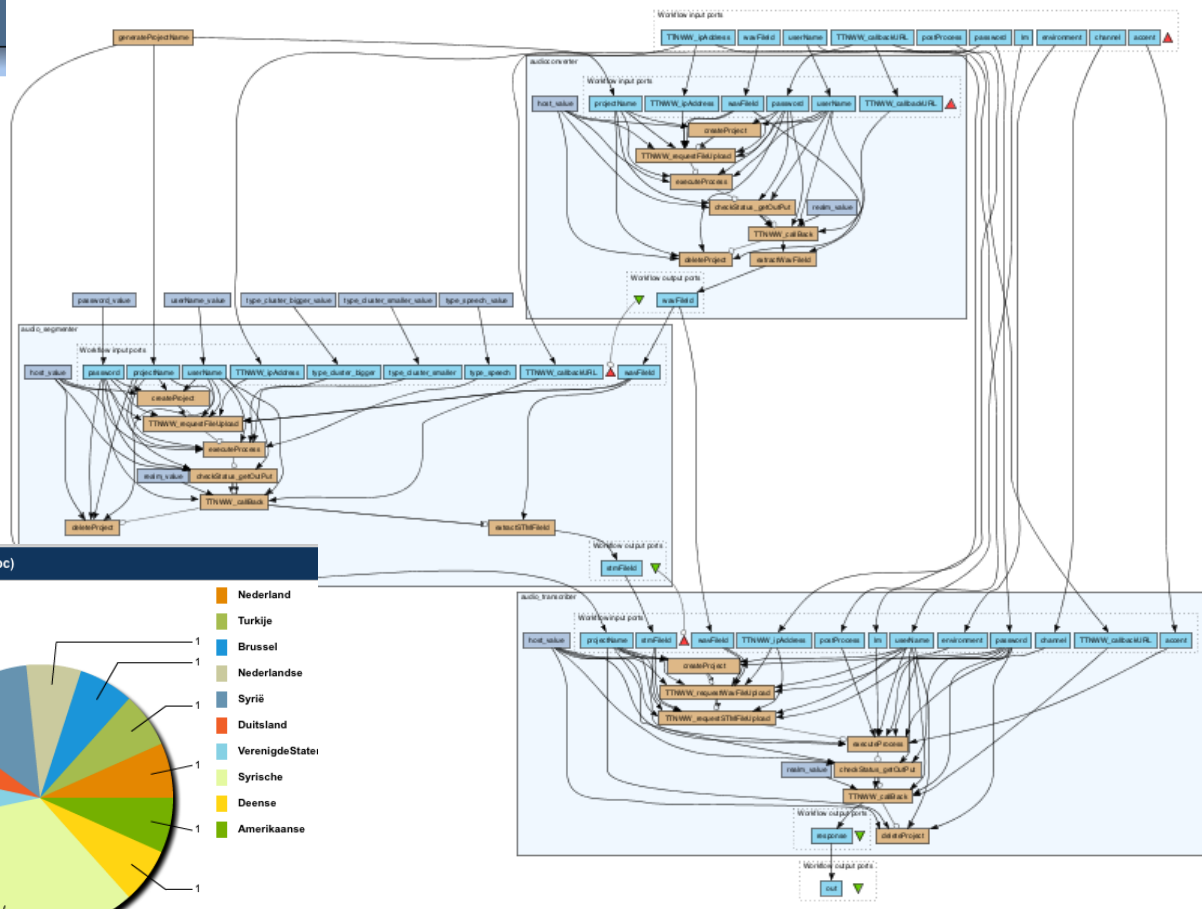


# EXISTING COMPONENTS

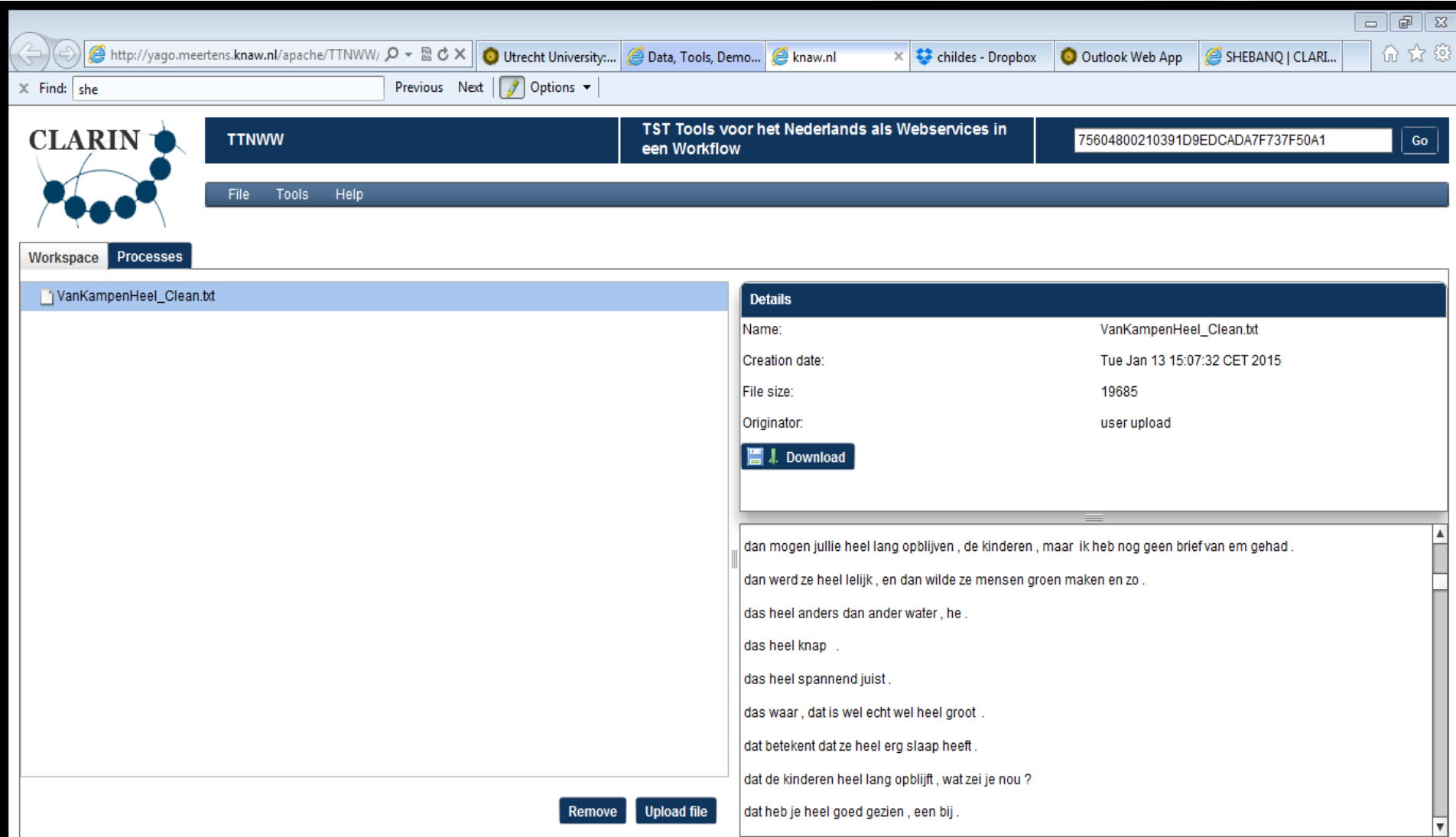
- For TTNWW, the [Taverna](#) workbench is used
- Taverna enables creation of one's own work flows, but
- In TTNWW only fixed, predefined work flows are available
  - Making your own work flow is pretty complicated
  - The targeted audience – humanities researchers – seldom want to do this



- Functionality
  - Data selection (upload/download)
  - Selection of processes
  - Calling processes
  - Puts results in a **temporary** work space
  - Only generic viewers







The screenshot shows a web browser window with the URL `http://yago.meertens.knaw.nl/apache/TTNWW/`. The browser tabs include 'Utrecht University...', 'Data, Tools, Demo...', 'knaw.nl', 'childes - Dropbox', 'Outlook Web App', and 'SHEBANQ | CLARIN...'. The search bar contains 'she'. The page header features the CLARIN logo, the title 'TTNWW', and a subtitle 'TST Tools voor het Nederlands als Webservices in een Workflow'. A search box contains the ID '75604800210391D9EDCADA7F737F50A1' with a 'Go' button. Below the header is a navigation menu with 'File', 'Tools', and 'Help'. The main content area has two tabs: 'Workspace' and 'Processes'. Under 'Processes', a file named 'VanKampenHeel\_Clean.txt' is selected. To the right, a 'Details' panel shows the following information:

Name:	VanKampenHeel_Clean.txt
Creation date:	Tue Jan 13 15:07:32 CET 2015
File size:	19685
Originator:	user upload

Below the details is a 'Download' button. The main content area displays the text of the file:

dan mogen jullie heel lang opblijven , de kinderen , maar ik heb nog geen brief van em gehad .  
dan werd ze heel lelijk , en dan wilde ze mensen groen maken en zo .  
das heel anders dan ander water , he .  
das heel knap .  
das heel spannend juist .  
das waar , dat is wel echt wel heel groot .  
dat betekent dat ze heel erg slaap heeft .  
dat de kinderen heel lang opblijft , wat zei je nou ?  
dat heb je heel goed gezien , een bij .

At the bottom of the workspace, there are 'Remove' and 'Upload file' buttons.



TTNWW

TST Tools voor het Nederlands als Webservices in een Workflow

75604800210391D9EDCADA7F737F50A1

Go

File Tools Help

- Text
  - Spelling variation
  - Tagging
- Audio
  - Parsing
  - Named Entity Recognition
  - Semantische rollen
  - Coreference

Workspace Processes

VanKampenHeel\_Clean.txt

### Details

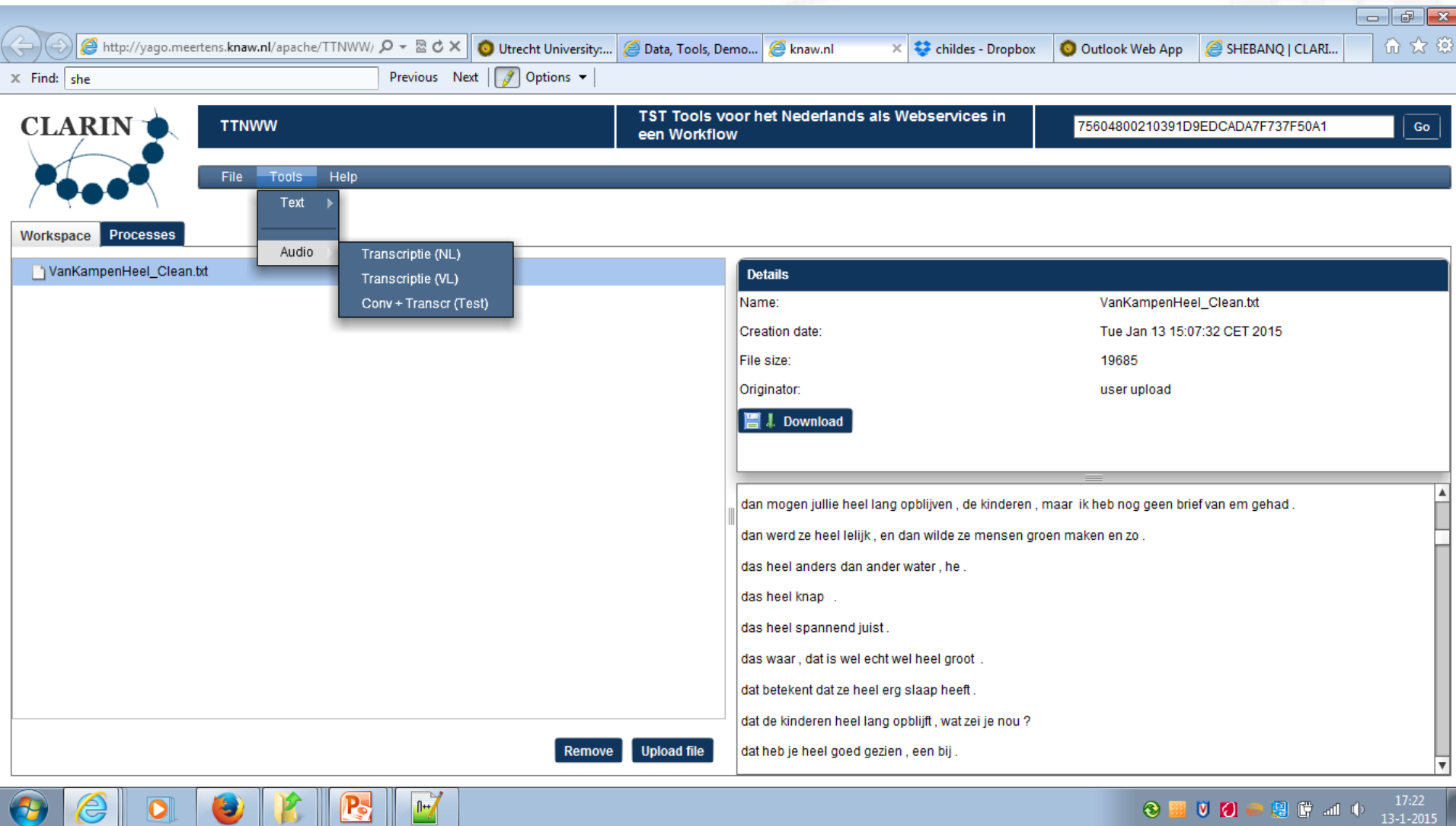
Name:	VanKampenHeel_Clean.txt
Creation date:	Tue Jan 13 15:07:32 CET 2015
File size:	19685
Originator:	user upload

 Download

dan mogen jullie heel lang opblijven , de kinderen , maar ik heb nog geen brief van em gehad .  
dan werd ze heel lelijk , en dan wilde ze mensen groen maken en zo .  
das heel anders dan ander water , he .  
das heel knap .  
das heel spannend juist .  
das waar , dat is wel echt wel heel groot .  
dat betekent dat ze heel erg slaap heeft .  
dat de kinderen heel lang opblijft , wat zei je nou ?  
dat heb je heel goed gezien , een bij .

Remove

Upload file



The screenshot shows a web browser window with the URL `http://yago.meertens.knaw.nl/apache/TTNWW/`. The page title is "TTNWW" and the subtitle is "TST Tools voor het Nederlands als Webservices in een Workflow". A search bar contains the text "75604800210391D9EDCADA7F737F50A1".

The interface includes a navigation menu with "File", "Tools", and "Help". The "Tools" menu is open, showing options for "Text" and "Audio". The "Audio" option is selected, and a sub-menu is visible with the following items:

- Transcriptie (NL)
- Transcriptie (VL)
- Conv + Transcr (Test)

The main workspace displays a file named "VanKampenHeel\_Clean.txt". Below the file list, there are "Remove" and "Upload file" buttons. On the right side, a "Details" panel provides information about the file:

Details	
Name:	VanKampenHeel_Clean.txt
Creation date:	Tue Jan 13 15:07:32 CET 2015
File size:	19685
Originator:	user upload
<a href="#">Download</a>	

Below the details panel, a text area displays the following Dutch text:

dan mogen jullie heel lang opblijven , de kinderen , maar ik heb nog geen brief van em gehad .  
 dan werd ze heel lelijk , en dan wilde ze mensen groen maken en zo .  
 das heel anders dan ander water , he .  
 das heel knap .  
 das heel spannend juist .  
 das waar , dat is wel echt wel heel groot .  
 dat betekent dat ze heel erg slaap heeft .  
 dat de kinderen heel lang opblijft , wat zei je nou ?  
 dat heb je heel goed gezien , een bij .

At the bottom of the interface, there are "Remove" and "Upload file" buttons. The Windows taskbar at the bottom shows the system clock at 17:22 on 13-1-2015.

# OVERVIEW

- Enrich your own resources (TTNWW)
- **Enrich your own resources (other applications)**
- Search in your own enriched resources (AutoSearch)
- Search in your own enriched resources (PaQu)
- Conclusions

# OTHER APPROACHES IN THE '*CLARIN*' DOMAIN

- [Weblicht](#) (login with your university account)
  - *Easy Mode* lets you choose pre-defined processing chains
  - *Features Mode* allows you to select individual features, and then choose from a selection of processing chains.
  - *Advanced Mode* allows you to build customized tool chains.
- Multiple languages, mainly German


## Sign in to Clarin EU Service Provider

Select your Identity Provider

[View Tool List](#) [HELPDESK](#)


If you cannot find your institution in the list above please select the "Clarin.eu website account" and use your credentials of the CLARIN website. For questions please contact [webmaster@clarin.eu](mailto:webmaster@clarin.eu).

CLARIN-D

← Clarin.eu website account  
 European Union

RUG  
 Netherlands

Universiteit Leiden  
 Netherlands

Graafschap College  
 Netherlands

TERENA Secretariat  
 Netherlands

Universiteit van Amsterdam  
 Netherlands

Tilburg University  
 Netherlands

Vrije Universiteit  
 Netherlands

NIOO (KNAW)  
 Netherlands

SURFnet bv  
 Netherlands



**Shibboleth.**

<https://weblicht.sfs.uni-tuebingen.de/WebLicht-4/>

or search for a provider, such as Example University.

Main Page **Chain 2** x +

Show tools with status:  dev  development  production  withdrawn

Next Choices (Double-click on an icon to add it to the chain)

<b>IMS</b> Morphology morphology i	<b>SfS</b> Convert to Negra Document Type: NEGRA Form: i	<b>Berlin-Brandenburg</b> Person Nar Named Entities: person i	<b>SfS</b> Berkeley Parser - Berke Parsing: tuebadztb i	<b>Berlin-Brandenburg</b> Lemmas2L Language: German Document Type: Lexicon Form: TCF Version: 0.4 entries.type: lemmas i	<b>Berlin-Brandenburg</b> Tokens2Le: Language: German Document Type: Lexicon Form: TCF Version: 0.4 entries.type: types i	<b>IMS</b> Constituent Parser Parsing: Tiger Treebank Tagset i	<b>Berlin-Brandenburg</b> CAB orthog orthography i	<b>SfS</b> German Named Entity F Model: conll2003 i
--	--	---	---	---	--	--	--	---

Input and Chain Selection

[Run Tools](#) [Clear Results](#) [Download chain](#)

<b>My Input</b> Plain Text Es war einmal ein Mädchen so schön das die Himmel weinte. ↕	<b>SfS</b> To TCF Converter Language: German Document Type: TCF TCF Version: 0.4 Text i X	<b>Berlin-Brandenburg</b> Tokenizer a Sentences Tokens i X	<b>IMS</b> Stuttgart Dependency F Part of Speech: STTS Tagset Parsing (Dep): No Empty Token Lemmas Parsing (Dep): tiger Parsing (Dep): false i X
--	--	---	--

### Available Annotations for: German Plain Text

- POS Tags/Lemmas
- Morphology
- Constituent Parses
- Dependency Parses
- Named Entities

### German Named Entity Recognizer

Annotation Layers:

Download as Excel sheet Download as CSV

- Simple view
  - text
  - sentences
  - Table view
    - tokens
    - POSTags
    - lemmas
    - namedEntities
  - Highlighted view
    - namedEntities

token ID	tokens	POSTags	lemmas	namedEntities
t1	Es	PPER	es	
t2	war	VAFIN	sein	
t3	einmal	ADV	einmal	
t4	ein	ART	ein	
t5	Mädchen	NN	Mädchen	
t6	so	ADV	so	
t7	schön	ADJD	schön	

Download TCF

### Input and Chain Selection

[Run Tools](#) [Clear Results](#) [Download chain](#)

**My Input** Plain Text

Es war einmal ein Mädchen so schön das die Himmel weinte.

**SfS To TCF Converter**

Language: German  
Document Type: TCF  
TCF Version: 0.4  
Text

**IMS Tokenizer**

Sentences  
Tokens

**IMS TreeTagger**

Part of Speech: STTS Tagset  
Lemmas

**SfS German Named Entity F**

Named Entities: tuebadz8

Done running tools.



Available Annotations for: German Plain Text

- POS Tags/Lemmas
- Morphology
- Constituent Parses
- Dependency Parses
- Named Entities

German Named Entity Recognizer

Annotation Layers:

- Simple view
  - text
  - sentences
  - Table view
    - tokens
    - POStags
    - lemmas
    - namedEntities
  - Highlighted view
    - namedEntities

language = de

[Download as Excel sheet](#)

[Download as CSV](#)

[ 1 - 30 / 46 ]

token ID	tokens	POStags	lemmas	namedEntities
t1	Israelische	ADJA	israelisch	
t2	Kampfflugzeuge	NN	Kampfflugzeug	
t3	haben	VAFIN	haben	
t4	in	APPR	in	
t5	der	ART	d	
t6	Nacht	NN	Nacht	
t7	neun	CARD	neun	

[Download TCF](#)

Input and Chain Selection

[Run Tools](#) [Clear Results](#) [Download chain](#)

<p><b>My Input</b> Plain Text</p> <p>Israelische Kampfflugzeuge haben in der Nacht neun Stellungen der syrischen Armee beschossen. Die Armee teilte mit, es handele sich um eine Antwort auf den von</p>	<p><b>Sfs</b> To TCF Converter</p> <p>Language: German Document Type: TCF TCF-Version: 0.4 Text</p>	<p><b>IMS</b> Tokenizer</p> <p>Sentences Tokens</p>	<p><b>IMS</b> TreeTagger</p> <p>Part of Speech: STTS Tagset Lemmas</p>	<p><b>Sfs</b> German Named Entity F</p> <p>Named Entities: tuebadz8</p>
--	---	---	--	---

Done running tools.

# OTHER APPROACHES IN THE '*CLARIN*' DOMAIN

- Panacea
  - Taverna based workflow system
  - You can use find workflows
  - You can add your own work flows
  - You can use workflows
    - In the Taverna workbench (that you must download and install)
  - Multiple languages, incl Italian and Spanish
  - Uses Freeling (see its on-line demo)

## Welcome to PANACEA

PANACEA, a STREP Project under EU-FP7, has developed a factory of Language Resources (LRs) in the form of a production line that automates all steps involved in the acquisition, production, maintenance and updating of the LRs required by Machine Translation and other Language Technologies.

The factory is a Web Service-based platform that integrates advanced technological components for:

- Monolingual and Parallel Text Acquisition and Pre-Processing
- Parallel corpora Alignment
- Bilingual Dictionary Production
- Monolingual Rich Information Lexica Production

Click [HERE](#) to learn how to begin using PANACEA.

Click [HERE](#) to get the acquired resources, i.e. the datasets produced during the project:

- Monolingual Corpora (raw text, n-grams, parsed, for EL, EN, ES, FR and IT)
- Parallel Corpora (sentence aligned, EN-EL and EN-FR)
- Monolingual Lexica (Verbal Subcategorization for EN, ES and IT, Noun Lexical-Semantic Classes, for EN and ES; MultiWords for IT)
- Bilingual Glossaries (EL-EN, FR-EN, DE-EN)

## PANACEA Web Portals



## News

**MT Summit 2013 PANACEA tutorial now available**

18  
Sep  
2013

[Read more](#)

**MT SUMMIT 2013 PANACEA TUTORIAL**

04  
Mar  
2013

PANACEA will hold a tutorial at the [Machine Translation Summit XIV](#) in Nice, France on 2 - 6 September 2013.

[Read more](#)

**8 PANACEA Papers accepted at COLING 2012**

23  
Nov  
2012

List of Accepted papers at the forthcoming 24th International Conference on Computational Linguistics COLING 2012, December Mumbai, India.

[Read more](#)



## Latest Services Available

TaaS API 2014-06-06, 07:15 am  
 Test service 2013-10-09, 03:34 am  
 corpus\_analysis 2013-04-16, 05:57 am  
 tpe\_subcat\_inductive 2013-04-15, 04:27 am  
 graf\_resource\_header\_generator 2013-04-03, 06:56 am

Workflows 

Home » Workflows

## Workflows

Search filter terms

Sort by: Rank

Showing 2 results. Use the filters on the left and the search box below to refine the results.

User: Marta Villegas

[Remove all filters](#)

Filter by type

Taverna 2 2

Filter by tag

- directory 2
- download 2
- freeling 2
- upload 2
- dependency 1
- graf 1
- tagging 1

Filter by user

- Marcpoch 33
- Muntsa Padró 13
- Valeria Quochi 8
- atoral 7
- Laura Rimell 3
- Thurmair 3
- Francesco Ru... 2
- Marta Villegas 2
- Prokopis 2
- Panacea Admin 1

Filter by licence

by-sa 2

**Taverna 2** **Freeling dependency for plain text data with input upload and output download (v1)** [View](#)  
[Download \(v1\)](#)

**Original Uploader**



Marta Villegas



This is the dependency workflow for plain text files in local directory using Freeling Graf. Freeling is run using the "keeptags" option to remove boilerplate and to keep paragraph tags info from the input data. The output is converted to graf format using the grafconverter\_postagging service. This workflow uses the download\_url processor for users who want to download the output files during the workflow execution and want the output files to follow the name of the input file. Exam...

Rating: 0.0 / 5 (0 ratings) | Versions: 1 | Reviews: 0 | Comments: 0 | Citations: 0

Viewed: 13 times | Downloaded: 7 times

Tags (6):

[dependency](#) | [directory](#) | [download](#) | [freeling](#) | [graf](#) | [upload](#)

### New/Upload

Workflow

### Log in / Register

Username or Email:

Password:

Remember me:

Need an account?  
[Click here to register](#)

[Forgot Password?](#)

### Popular Tags

25 tags

[\[All Tags\]](#)

[basicxcex](#) | [bilingual](#) | [cleaner](#) | [cqp](#) | [crawled](#) | [dependency](#) | [dear](#) | [directory](#) | [download](#) | [english](#) | [example](#) | [freeling](#) | [graf](#) | [hunalign](#) | [ilsp](#) | [lexical acquisition](#) | [lists](#) | [merge](#) | [noun classification](#) | [panacea](#) | [parser](#) | [pos tagging](#) | [sentence alignment](#) | [spanish](#) | [tagging](#)



# OTHER NLP SERVICES IN THE '*CLARIN*' DOMAIN

- Several services at LINDAT/CLARIN (Czech Republic), including Treex::Web (Czech, English)
- Several on-line services at CLARIN-DK, multiple languages
- Some services at CLARIN-PL (mainly for Polish)
- Language Analysis Portal by CLARINO (under development, mainly Norwegian)
- ...

# OVERVIEW

- Enrich your own resources (TTNWW)
- Enrich your own resources (other applications)
- **Search in your own enriched resources (AutoSearch)**
- Search in your own enriched resources (PaQu)
- Conclusions

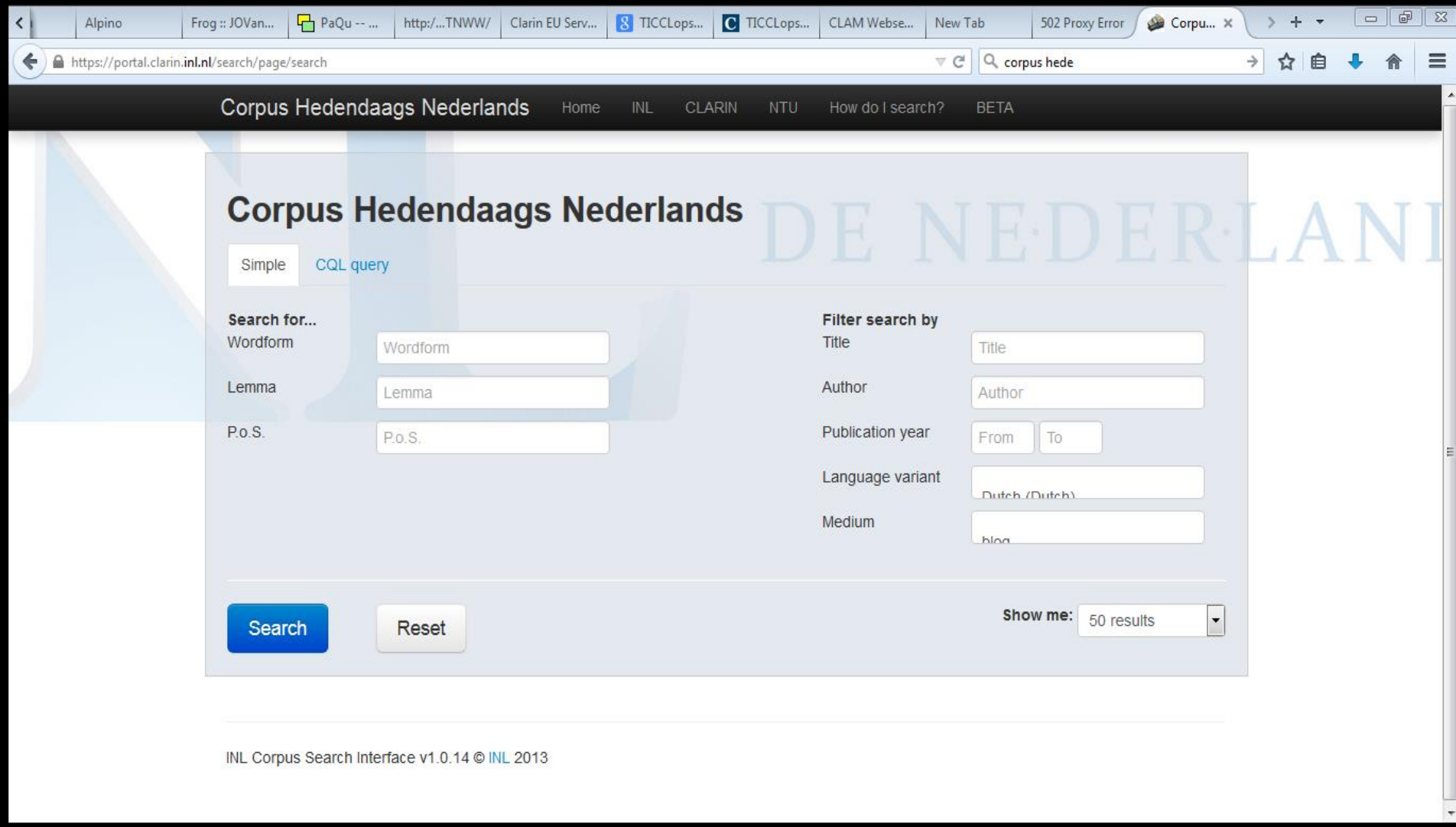
# SEARCH & ANALYSIS

- Enriching your data is fine but
- What will you do with a (large and complex) set of XML files?
- You need options to search in the data and to analyze the data
- Two projects running in NL
  - AutoSearch (INL)
  - PaQu (Groningen University)

# AUTOSEARCH

- (under development, available March 2015)
- For pos-tagged data
- Supported formats:
  - [FoLia](#) (Frog yields Folia format)
  - [TEI](#)
- Search Interface: [Corpus Hedendaags Nederlands](#) (Corpus of Modern Dutch) interface



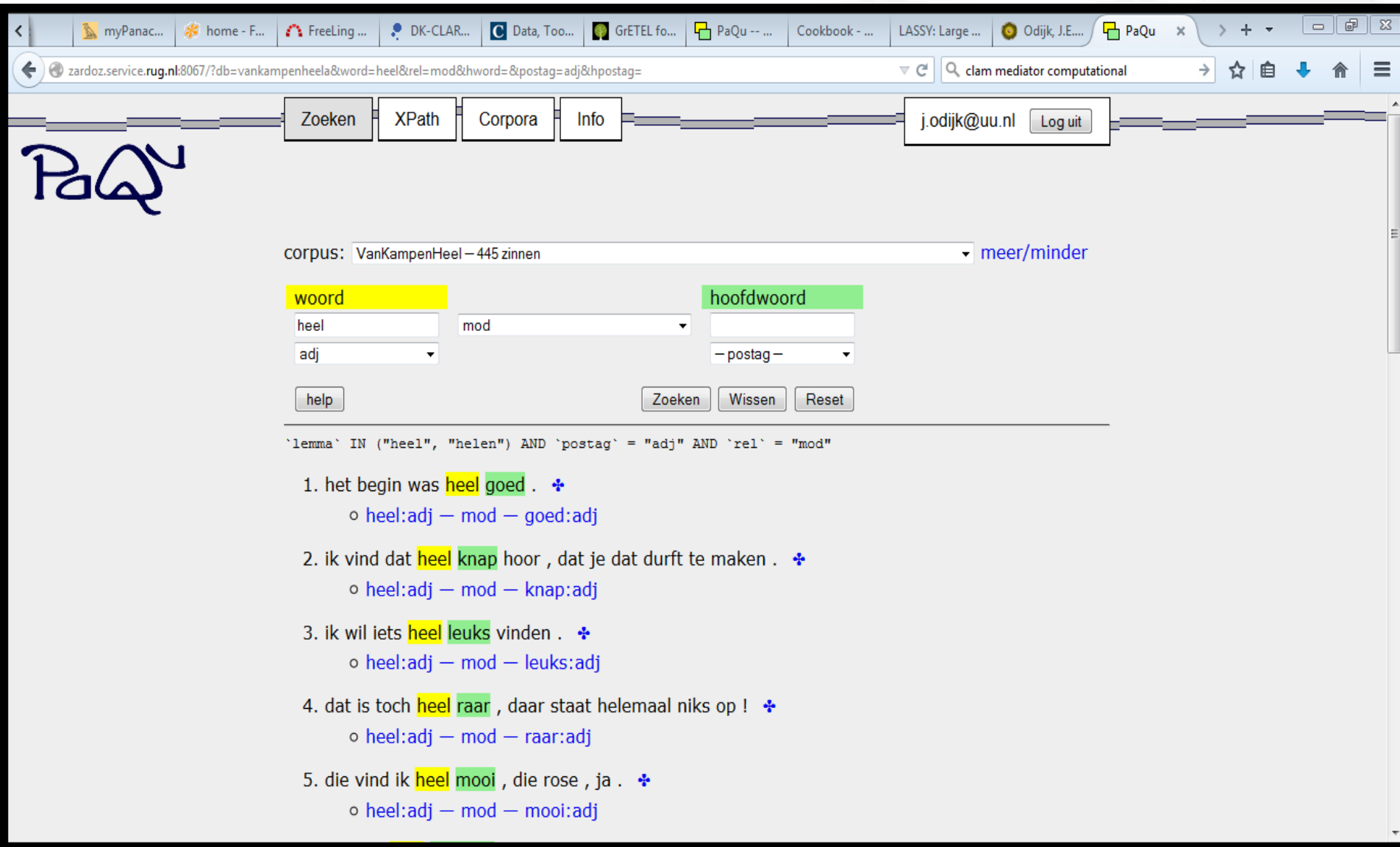
A screenshot of a web browser displaying the search interface for the Corpus Hedendaags Nederlands. The browser's address bar shows the URL "https://portal.clarin.inl.nl/search/page/search" and the search term "corpus hede". The page header includes navigation links: "Home", "INL", "CLARIN", "NTU", "How do I search?", and "BETA". The main content area is titled "Corpus Hedendaags Nederlands" and features two search modes: "Simple" (selected) and "CQL query". Under "Search for...", there are input fields for "Wordform", "Lemma", and "P.o.S.". Under "Filter search by", there are input fields for "Title", "Author", "Publication year" (with "From" and "To" sub-fields), "Language variant" (with a dropdown menu showing "Dutch (Dutch)"), and "Medium" (with a dropdown menu showing "blog"). At the bottom of the search area, there is a blue "Search" button, a grey "Reset" button, and a "Show me:" dropdown menu set to "50 results". The footer of the page reads "INL Corpus Search Interface v1.0.14 © INL 2013".

# OVERVIEW

- Enrich your own resources (TTNWW)
- Enrich your own resources (other applications)
- Search in your own enriched resources (AutoSearch)
- **Search in your own enriched resources (PaQu)**
- Conclusions

- under development
  - available July 2015
  - Preliminary version available now!
- For fully parsed data
- Supported formats:
  - [LASSY XML](#) (yielded by TTNWW)
  - Plain text (and PaQu does the parsing)
- Search Interface: [Groningen Word Relations Interface](#)

- Small Experiment:
  - I selected the child-directed sentences containing *heel, erg, zeer* from the CHILDES Van Kampen corpus
  - Cleaned the text from annotations by a script
  - Loaded them to PaQu, and did an analysis.
  - Made a comparison with my manual annotation (see third lecture on the first day)- for the *mod* relation only.
  - Extended the manual annotation to cover multiple instances in one sentence (445 -> 452)
  - After comparison, I corrected the errors in my manual annotation (*revised manual annotation*)



zardoz.service.rug.nl:8067/?db=vankampenheela&word=heel&rel=mod&hword=&postag=adj&hpostag=

clam mediator computational

Zoeken XPath Corpora Info j.odijk@uu.nl Log uit

PaQu

corpus: VanKampenHeel – 445 zinnen [meer/minder](#)

woord: heel mod hoofdwoord:   
 adj - postag -

help Zoeken Wissen Reset

`lemma` IN ("heel", "helen") AND `postag` = "adj" AND `rel` = "mod"

1. het begin was heel goed . ✦  
 o heel:adj – mod – goed:adj
2. ik vind dat heel knap hoor , dat je dat durft te maken . ✦  
 o heel:adj – mod – knap:adj
3. ik wil iets heel leuks vinden . ✦  
 o heel:adj – mod – leuks:adj
4. dat is toch heel raar , daar staat helemaal niks op ! ✦  
 o heel:adj – mod – raar:adj
5. die vind ik heel mooi , die rose , ja . ✦  
 o heel:adj – mod – mooi:adj

zardoz.service.rug.nl:8067/?db=vankampenheela&word=heel&rel=mod&hword=&postag=adj&hpostag=

o [heel:adj](#) — [mod](#) — [netjes:adj](#)

[vorige](#) | [volgende](#)

---

tijd: 19ms

[tellingen — algemeen](#)

Selecteer twee of meer elementen om ze te koppelen:

**woord**       **hoofdwoord**

lemma     relatie     lemma

postag     postag

[tellingen van combinaties](#)

```
`lemma` IN ("heel", "helen") AND `postag` = "adj" AND `rel` = "mod"
```

	lemma	rel	hpostag
394	heel	mod	adj
23	heel	mod	vnw
18	heel	mod	bw
14	heel	mod	n
2	heel	mod	ww
1	heel	mod	vz

tijd: 14ms

[download](#)

mede mogelijk gemaakt door:

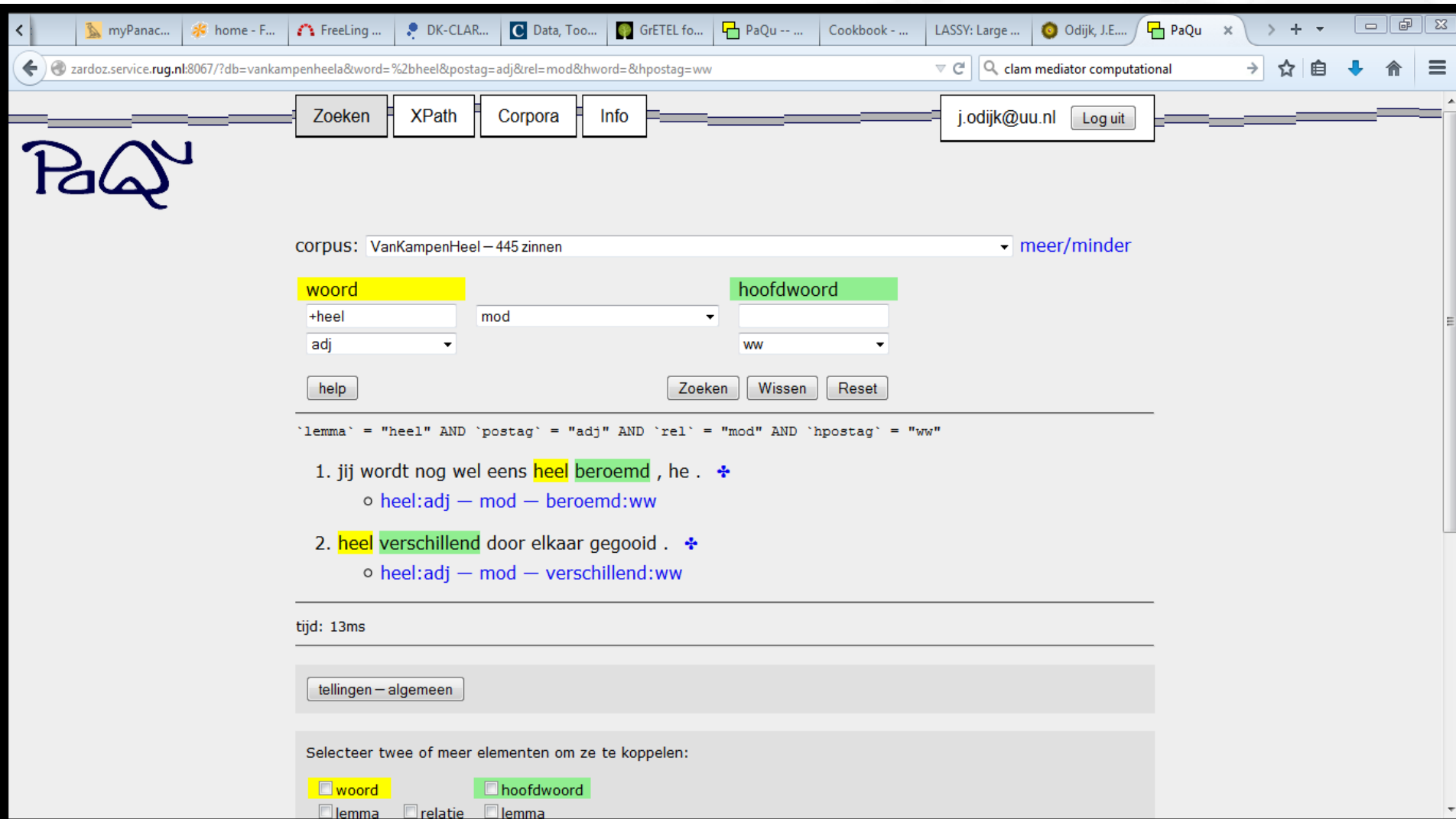
- Comparison Alpino v. Manual v. Revised:

Case	Alpino	Manual	Revised
Mod adj	394	428	426
Mod vnw	23	--	--
Mod bw	18	--	--
Mod n	14	10	10
Mod ww	2	2	1
Mod vz	1	0	0
Predet n	0	1	2
predicative	0	7	8
unclear	0	4	5

- Comparison Alpino v. Manual:

Count of examples Column Labels							
Row Labels	Amod	Nmod	NpreDet	pred	unclear	Vmod	Grand Total
Amod	387	1		6	4		398
Bwmod	16					1	17
Nmod	6	6	1				13
Vmod	2						2
VNWmod	17	3		1			21
Vzmod						1	1
<b>Grand Total</b>	<b>428</b>	<b>10</b>	<b>1</b>	<b>7</b>	<b>4</b>	<b>2</b>	<b>452</b>





[Zoeken](#) [XPath](#) [Corpora](#) [Info](#)
j.odijk@uu.nl [Log uit](#)

corpus:  [meer/minder](#)

woord hoofdwoord

---

``lemma` = "heel" AND `postag` = "adj" AND `rel` = "mod" AND `hpostag` = "ww"`

1. jij wordt nog wel eens heel beroemd , he . ✚
  - o [heel:adj](#) – [mod](#) – [beroemd:ww](#)
2. heel verschillend door elkaar gegoid . ✚
  - o [heel:adj](#) – [mod](#) – [verschillend:ww](#)

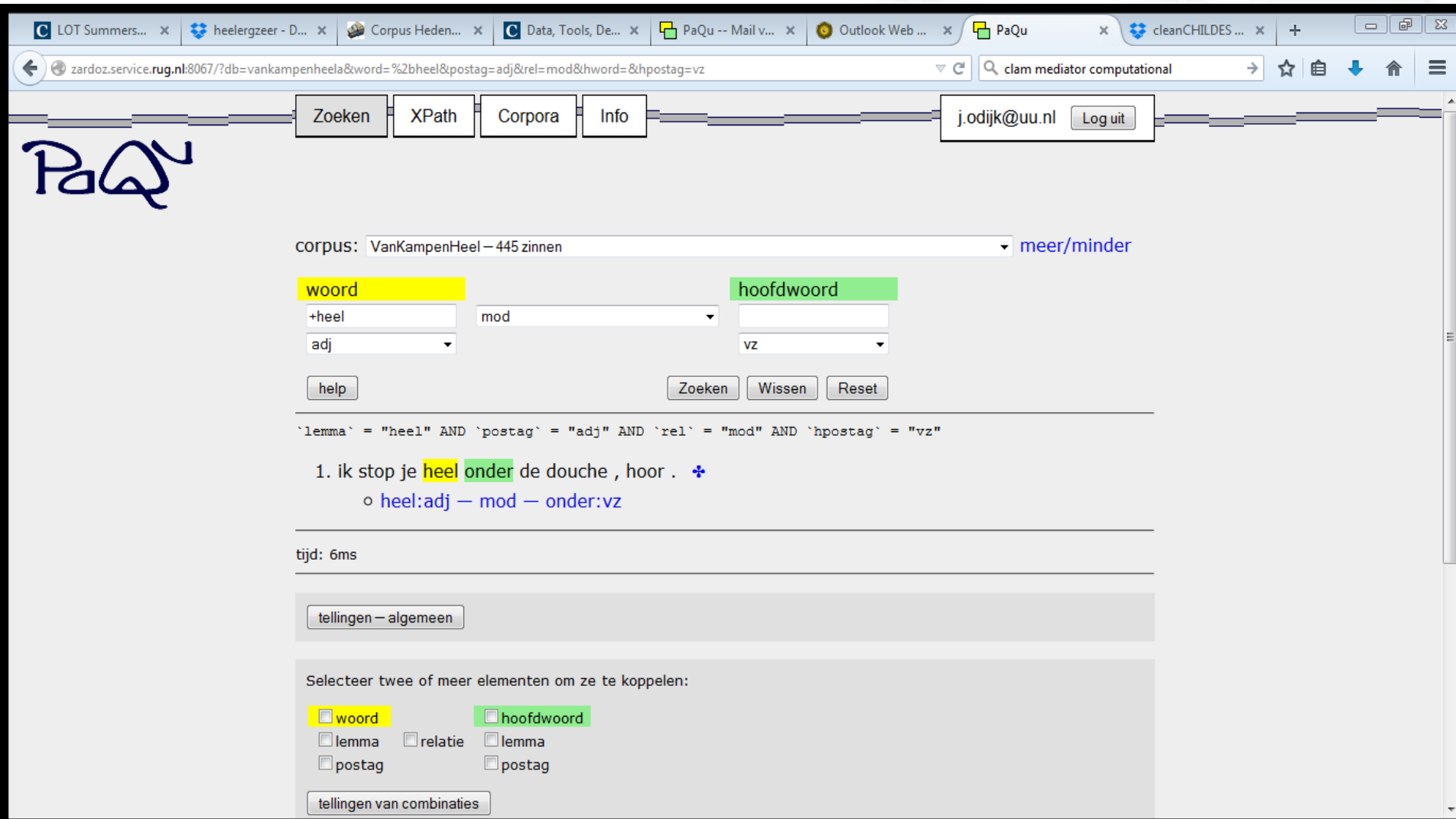
---

tijd: 13ms

Selecteer twee of meer elementen om ze te koppelen:

woord  hoofdwoord

lemma  relatie  lemma



zardoz.service.rug.nl:8067/?db=vankampenheela&word=%2bheel&postag=adj&rel=mod&hword=&hpostag=vz

clam mediator computational

Zoeken XPath Corpora Info j.odijk@uu.nl Log uit

PaQU

corpus: VanKampenHeel – 445 zinnen [meer/minder](#)

**woord** **hoofdwoord**

+heel mod   
 adj vz

help Zoeken Wissen Reset

`lemma` = "heel" AND `postag` = "adj" AND `rel` = "mod" AND `hpostag` = "vz"

1. ik stop je **heel** **onder** de douche , hoor . ✦  
 o heel:adj – mod – onder:vz

tijd: 6ms

tellingen – algemeen

Selecteer twee of meer elementen om ze te koppelen:

**woord**  **hoofdwoord**  
 lemma  relatie  lemma  
 postag  postag

tellingen van combinaties

- Accuracy Alpino:
  - TolAlpino = VNWmod and BWmod counted as Amod

Alpino v Manual	0,87
Alpino v Revised	0,88
TolAlpino v manual	0,94
TolAlpino v revised	0,96

- Initial Experiment is promising
  - Automatic parsing gives reasonable results for a small experiment involving CHILDES data
    - But it concerns adult speech
    - It mostly concerns a very local grammatical relation
- Many extensions are still needed
  - CHILDES CHAT format (and many other formats) should be natively supported, including cleaning
  - Search and analysis of data and all metadata should be supported

# OTHER SYSTEMS

- I am not aware of any, except perhaps:
  - [TüNDRA](#) - the *Tübingen aNnotated Data Retrieval Application*

# OVERVIEW

- Enrich your own resources (TTNWW)
- Enrich your own resources (other applications)
- Search in your own enriched resources (AutoSearch)
- Search in your own enriched resources (PaQu)

## ➤ **Conclusions**

# CONCLUSIONS

- TTNWW enables linguists to enrich their own data (for Dutch)
- There is an increasing number of such applications for other languages, in large part thanks to CLARIN
- Applications to upload such data into search engines are under development
- Initial results of a small experiment are promising

Thanks for your Attention!